Using GOMS to Model Individual Differences in a Competence Assessment Task

Hadeel Ismail (hi71@sussex.ac.uk)
Department of Informatics, University of Sussex
Brighton, BN1 9QJ, UK

Peter C-H. Cheng (p.c.h.cheng@sussex.ac.uk)
Department of Informatics, University of Sussex
Brighton, BN1 9QJ, UK

Abstract

This study aims at modelling individual differences using GOMS. In an attempt to evaluate a competence assessment task in natural language, results revealed limitations of a previous GOMS model that was used to design the task (Ismail & Cheng, 2021). The task, Chunk Assessment by Stimulus Matching (CASM), exploits measurements of chunk signals to assess competence in the English language. It was tested with 34 speakers of English as a second language. Results were compared against the initial GOMS models. The models' predictions were partially supported, showing substantial performance differences between the levels of expertise. Contrary to expectations, major differences were found amongst those at the same level of expertise. A refinement of the models was built to coherently capture differences between and within levels of competence.

Keywords: chunking; GOMS; individual differences; language competence; model evaluation; pause analysis

Introduction

In cognitive science, pauses in recall and copying tasks have long been associated with mental processes. Some studies have specifically examined how pauses might reflect aspects of an expert's memory and inspired others to examine how these temporal measures might distinguish between experts and novices in specific domains. The classic studies performed by Chase and Simon (1973) involved memory and perceptual tasks for replicating item positions on a chessboard. Their findings reveal that expert's ability to remember far more positions than novices with close to perfect replications in memory tasks, and returning less to view the stimulus during copying tasks. Their observed physical actions are typically explained by the chunking theory (Cowan, 2001; Gobet et al., 2001; Miller, 1956). In short, the theory clarifies that during the process of perceiving domain-specific information, the amount held in working memory (WM) is dependent on the individual's representation of related information in their long-term memory. The more knowledgeable a person is, the richer their representation, which in turn assists in perceiving large chunks of meaningful information. Therefore, an expert's knowledge overcomes the limitations of WM providing them with the advantage of having larger chunks that encode more units of information than a novice.

Chunking theory informed the studies of Cheng and colleagues. They observed individuals' hand transcriptions of mathematical equations (Cheng, 2014; Cheng & Rojas-Anaya, 2007), English sentences (Zulkifli, 2013), and programming scripts (Albehaijan & Cheng, 2019) using behavioral measures that included pauses, writing durations, and the number of times a stimulus is viewed. These micro-behavioral measures, captured at a millisecond timescale, show some potential as metrics for assessing competency.

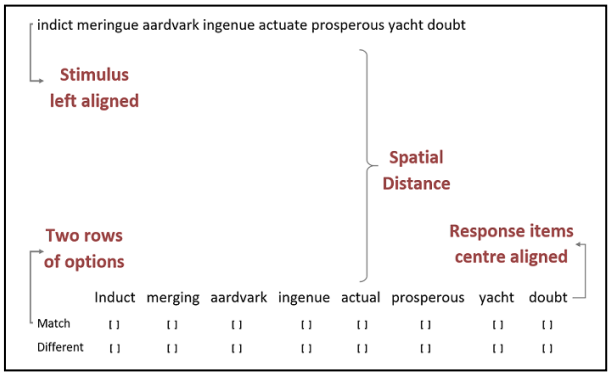

Figure 1: CASM task (Ismail & Cheng, 2021) – text in red are explanatory labels about the interface.

Rather than logging pen strokes, a recent paper proposed the *Competence Assessment by Stimulus Matching technique*, CASM, that utilizes the mouse device in word matching tasks (See Fig.1), in order to assess competence in the English language (Ismail & Cheng, 2021). The researchers used GOMS cognitive task analysis (Card, Moran, & Newell, 1983) to help in designing the tasks for CASM. In particular, GOMS was used to design tasks that would promote expert's use of the chunks, available from their superior knowledge, in order to maximize their performance compared to novices and thereby provide strong measures of competence. From the GOMS models of alternative tasks designs the ones with the largest theoretical differentiation across level of competence were adopted.

The empirical evaluation, to be summarized in the third section, shows substantial performance differences between the levels of expertise. Critically, it also revealed a limitation of the original GOMS models. Major differences were found amongst those at the same level of competence. The aim here is to build models that more coherently capture differences between and *within* levels of competence. In a sense, this study challenges the claim that GOMS does not take into account individual differences (Olson & Olson, 1990).

One motivation for this work is to continue developing CASM by controlling for sources of individual differences unrelated to competence, in order to improve the quality of

Table 1: The four designs of the CASM task

Stimulus-response matching	Stimuli presentation	
	Constant display, CD	Voluntary display, VD
Word to word, WW	CDWW	VDWW
Part-word to word, PW	CDPW	VDPW

the CASM measures. We are following Gong & Kieras's (1994) general advice to include GOMS modelling as part of our iterative cycle for system development. So, the refinement of the models to encompass a range of individual differences are examined in the fourth section.

Design of the CASM Task

CASM attempts to assess an individual competence in terms of the chunks they possess. The basic idea is to log an individual's' interaction as they decide if given stimuli words correctly match corresponding words in the response group (Fig.1). Individuals must quickly and accurately compare and click their responses. The task is designed in a way that encourages the use of chunking, this includes a separation between the stimuli and the responses and the non-alignment of the two which motivates participants' use of a strategy involving recognizing and remembering the words. It is assumed that depending on an individual's level of English competence, the number of words held in WM would be manifested in their behavior. An expert's prior knowledge would provide them with the advantage of quickly recognizing words and capturing multiple words into their WM, whereas a novice's limited knowledge would constrain their chunking thus forcing them to refer back to the stimuli more times than an expert. Such differences are reflected in the length of pauses preceding their clicks as well as their pause patterns.

Since the design space of the CASM task was large, GOMS analysis was applied to find tasks that maximize the differentiation between experts and novices (Ismail & Cheng, 2021). Two manipulations of stimuli visibility and two types of stimuli to response matching were proposed (see, Table 1). The model predicts that across all four tasks, expert's pauses would be shorter than a novice, with higher differentiation in *part-word to word* (PW) tasks compared to *word to word* (WW) tasks of the same presentation condition (Table 2, top).

In the *constant display* condition (CD) the stimulus and the response items remained visible throughout the trials. The duration of pauses between clicks reflects differences in chunking and hence is a potential measure of competence. In the *voluntary display* condition (VD), the stimulus and the response items were not simultaneously displayed. On loading the screen, the response items (bottom) were made visible with the stimulus (top) remaining concealed under an interactive grey box. A hover of the mouse pointer over the box will reveal the stimulus and mask the response items. Upon hovering away to mark their responses, the stimulus and the response will revert to their initial visibility states. Participants were allowed to hover over the stimulus as many times as they needed. With the VD condition, measures of chunking include the number of hovers made to view the stimulus and the duration of time spent clicking between views.

Table 2: Compares the original model's median of pauses prediction (Ismail & Cheng, 2021) to the observed group median of pauses for the five least and most competent

	Model predictions (msc)		
Task type	**Novice**	**Expert**	**Difference**
CDWW	2275	1070	1205
VDWW	3205	1070	2135
CDPW	3175	1120	2055
VDPW	4950	1120	3830
	Experimental results (msc)		
Task type	**Low**	**High**	**Difference**
CDWW	2971	1725	1246
VDWW	3946	1987	1959
CDPW	4632	2210	2422
VDPW	5430	2240	3190

These display conditions were combined with two matching conditions: *word to word* (WW) or *part-word to word* (PW) matching tasks. The WW condition consisted of matching whole words in the stimulus with whole words in the response (Fig.1). Since novices might not be familiar with many of the words presented, their basic strategy in matching WW is expected to consist of decomposing a word into parts that are separate chunks, thus filling their WM capacity more quickly than for experts. In PW condition, each word in the stimulus was broken up and presented as a string of syllables with equal spacing between them and the following word's syllables. For example, a stimulus containing the words "indict meringue aardvark" would be presented as "in dict me ringue aard vark" and these, as with the WW condition, were matched with complete words in the response. In this PW task it is assumed that an expert, at a slower pace than in WW, would still be able to recognize and chunk whole words. However, the PW task might encourage a novice to chunk one syllable at a time, thus switching many more times between the stimulus and the response prior to making a matching decision and clicking.

Empirical evaluation of CASM tasks

To assess the model predictions, an empirical evaluation of CASM tasks was carried out.

Method

The experiment is a within-subject design. It was approved by the University of Sussex Science School's ethics committee.

Participants. The participants were 34 adults, eight males, and twenty-six females, whose ages ranged between 18 to 54 years old. They were recruited on the basis that they spoke Arabic as their first language and English as a second, but with varying degrees of competence in English.

Materials. The experiment involved three stages. The first was a questionnaire that gathered participants' background, general ability, and confidence in using the English language.

The second was a generic vocabulary size test that assessed their overall level of competence. Scores gathered from the first two stages determined the participant's overall level of English language competence. The final stage was the CASM task which included the four conditions in Table 1. Each condition started with a set of instructions followed by three practice sessions and then twelve trials. Each trial consisted of eight words. The level of trial difficulty was determined by the frequency of the target words and their number of syllables. There were four types of word frequency, ranging from high to low, and three syllabic levels that included two, three, and four-syllable words. The order of the conditions received by the participants was counterbalanced, and the trials within were presented in random order.

Procedure. All of the materials were delivered online and were run on their personal computers using their own mouse. They had the option to complete all stages in one or up to three sittings. However, once a stage has been launched, it must be entirely completed without any interruptions. Specific instructions were given at the start of the CASM task which included matching the words as quickly and as accurately as possible and refraining from removing their hands from the mouse unless instructed otherwise.

In order to examine expert vs novice performance, participants were rank ordered according to their independent measure of competence. A systematic check was applied showing the top and bottom five individuals being reasonably consistent and thus were chosen to represent the extremes.

Experimental Results

To compare the performance of the five highest and lowest competent individuals (HC & LC), a group median was calculated using the mean pause of each participant. Each participant's mean pause was calculated from the median pause for each of their trials.

Across the four tasks substantial differences in the pauses exist (Table 2, bottom). This confirms the original model's prediction of pause lengths decreasing with increasing competence (Table 2, top). Moreover, PW tasks seem to have a higher differentiation effect compared to WW tasks of the same display conditions, in line with the predictions of the model (4th column in Table 2). Finally, the difference in pauses between HC and LC participants were comparable to the predictions of the model with an absolute difference of 20% or less (4th column in Table 2). According to HCI heuristics, an engineering model is acceptable if it reaches a level of accuracy of at least 80% (John & Kieras, 1994). Contrary to expectations, the absolute pause times were underestimated for both HC and LC individuals, with a level of accuracy as low as 51% (2rd & 3rd column in Table 2).

The divergence between human performance times and that predicted by the model indicate processes that the original model failed to foresee. This is likely due to variations in participants' strategies. Information concerning their patterns of clicks and hovers in the VD condition allowed us to carry a detailed examination of their strategies. The results show that there are intra-participant strategy differences, Fig. 2. The figure displays a selection of participants from both groups in the VDWW and VDPW conditions. Clear differences in terms of the number of hovers and pattern of clicks exist both at the level between and within groups.

In tasks involving WW matching, the basic assumption made by the original GOMS model is that novices would follow a *single-view-single-pick* strategy (Fig. 2, P48-A) which involves chunking one word during one hover/view of the stimuli, making a comparison and then clicking an answer (Ismail & Cheng, 2021). However, a *multi-view-single*-pick strategy was sometimes applied, where a single click is preceded by several hovers (e.g., Fig. 2, P48-B). One explanation for such behaviour is that LC individuals are uncertain of the chunked item in memory and go back to the stimuli for further verification prior to giving an answer.

In terms of PW matching, the original GOMS model assumed that since words were presented as parts, then novices might find it more convenient to follow a *multi-view-single-pick* strategy by chunking one syllable at a time, comparing each part of a word separately until reaching a decision (Ismail & Cheng, 2021). This would imply that in VDPW, the number of hovers made prior to clicking an answer would equal to the number of parts the word is divided into. The experimental results pertaining to the least competent

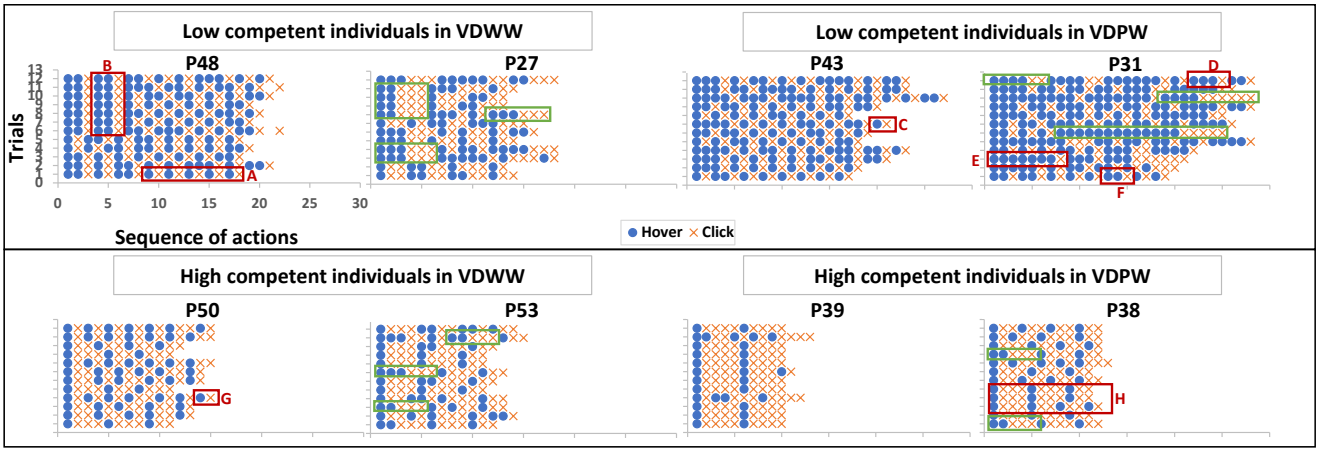

Figure 2: Various strategies applied by the participants across and within groups

individuals reveal far more complex strategies than what was originally assumed and they are:

1- *Single-view-single-pick* strategy; indicating their ability to group the parts of one word in one view of the stimuli (Fig. 2, P43-C).
2- *Multi-view-single-pick* strategy; conforming to the model's overall prediction but differing in terms of the number of hovers made (Fig. 2, P31-D, E, F). The number of hovers might either be less than, more than, or equal to the number of word parts. Such behavior may be explained in terms of one or a combination of the following:
    a. Inability to group all parts of a word in one view.
    b. Difficulty in locating word boundaries impacting their word recognition process
    c. Uncertainty of the word chunked in WM.

Experts were assumed to follow a *single-view-multi-pick* strategy by consistently loading large chunks of words into WM (e.g., P39 in Fig. 2). However, consistency varied, sometimes high competent (HC) participants would engage in a *single-view-single-pick* strategy, similar to that of a novice (Fig, 2, P50-G). Other times, they would apply an alternating strategy (Fig. 2, P38-H). Such variations could imply a lack of motivation in maximizing their use of WM.

There were instances where HC and LC participants would both engage in a *multi-view-multi-pick recoding* strategy, a purely strategic tactic that does not reflect competence (e.g., green highlights in Fig. 2). The initial GOMS models assumed that, across both extremes, once a word in WM is compared to the response item, the process is immediately followed by a mouse click (Ismail & Cheng, 2021). However, by applying a recoding strategy, participants might generate a list of decision codes in their memory by hovering over the stimuli, chunking a word or so, hovering away to reveal the responses, making a comparison, encoding their decision and then proceeding to process the next word in the stimuli without clicking an answer pertaining the first word(s). This would continue for a few times until enough codes have been loaded into WM, only then would they proceed to click multiple answers at once.

Although the experimental results support the predictions of the initial model (Ismail & Cheng, 2021) in terms of overall pause differences between the experts and the novices. Findings reveal that the strategies applied are far more complex than the previous model, thus allowing for individual differences within groups to arise.

GOMS Models

Based on the HC and LC individuals' performances, two models were generated that cohesively account for the different individual strategies that exist between and within the expert and novice groups (Fig. 3 and Fig. 4). The models represent the processes when working under the CD condition in PW and WW matching tasks. Overall, the models are divided into two parts, everything prior to the process "move eye to response area" concerns chunking processes, and everything there after deals with comparing, matching and mouse moving processes. The new models are more complex than the original attempt as they encompass individual differences at all levels. The green dashed lines in the figures point to WW processes, the purple dashed lines are associated with PW, while the black solid lines are those shared by both tasks. It is worth noting that the overall construct of the models under the CD condition is similar to the VD condition with the exception of having a hover over/away action whenever alternating views between the stimuli and the responses. The models explain the chunking process in terms of nested loops. The novice and the expert models differ in the number of loops and the type of processes contained within each loop (see the orange brackets in Fig. 3 and Fig. 4). This explains the inter-participant differences. Moreover, not all loops are experienced by all members of the same group, which explains for the intra-participant differences.

The first loops in both models concern the chunking process. In WW tasks novices break each word into its parts, individually processing them until a whole word is recognized and captured in WM (Fig. 3, NLP1(WW)). Experts have the ability to immediately recognize a word and capture it in memory, thus looping around ELP1(WW) (Fig. 4) as many times to generate a chunk of words. In PW tasks, the presentation of the words slows down the recognition process. Experts now experience two loops when chunking. The ELP1A(PW), shows how an expert must process enough syllables until a word is recognized. The second loop ELP1B(PW) explains the forming of a chunk of words. Novices on the other hand seem to experience much more difficulty, as their NLP1(PW) loop is more complex by including "the boundary confusion" decision process. Since the stimuli presents the words with equal spacing between all syllables, novices might find it difficult to distinguish word boundaries. With this added level of complexity, novices might not be able to chunk a whole word in one view causing them to return to the stimuli as many times as needed until a whole word is successfully recognized. Moreover, novices might be uncertain during the process of comparing the chunked item with the response word, and may wish to verify their answer prior to clicking. This in turn introduces the NLP2(Both) loop, that gives another explanation for their returns to the stimuli.

These differences across groups, in both WW and PW, show how novices experience increased cognitive effort in processing the presented words limiting their chunking ability, and causing them to perform multiple returns to the stimuli, therefore experiencing many long pauses between clicks. In contrast, experts have the opportunity to chunk more than one word per view, thus demonstrating shorter pauses between clicks (Table 2).

Within these two groups individual differences were found, which could be explained by the number of times individuals choose to go through the loops depicted in the models. For instance, in WW matching tasks, if a word is very familiar to a novice, then they might apply a *single-view-single-pick* strategy therefore bypassing the NLP2 loop in Fig.3. Otherwise, an unfamiliar word would produce a *multi-view-single-*

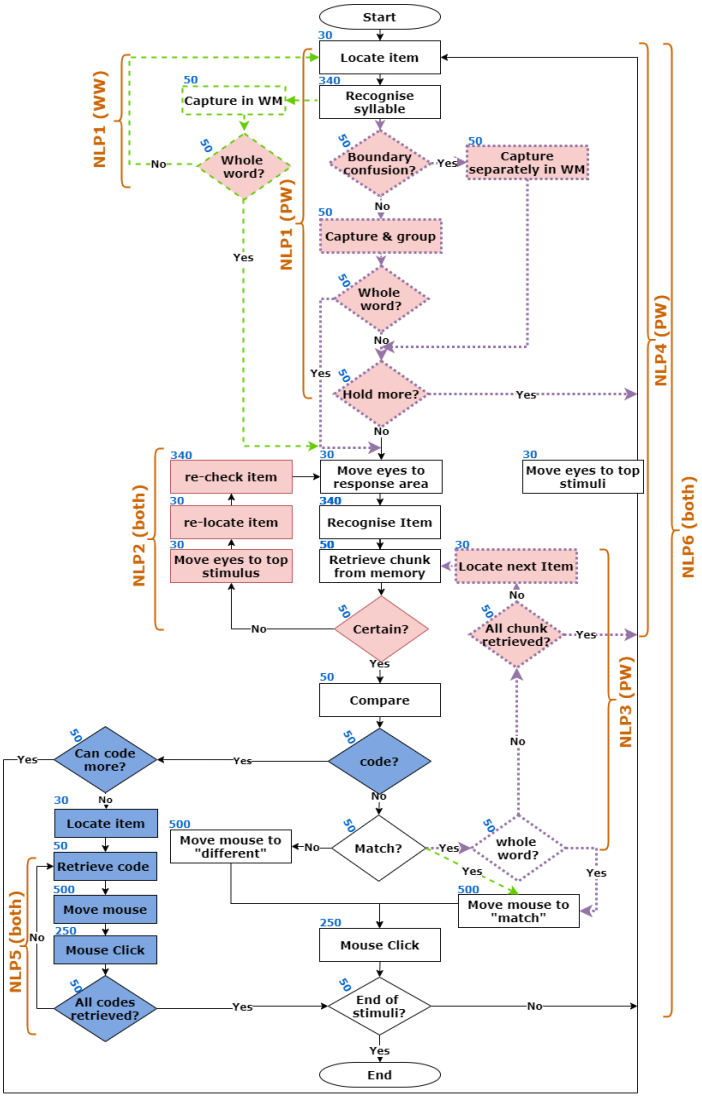

Figure 3: Novice Model

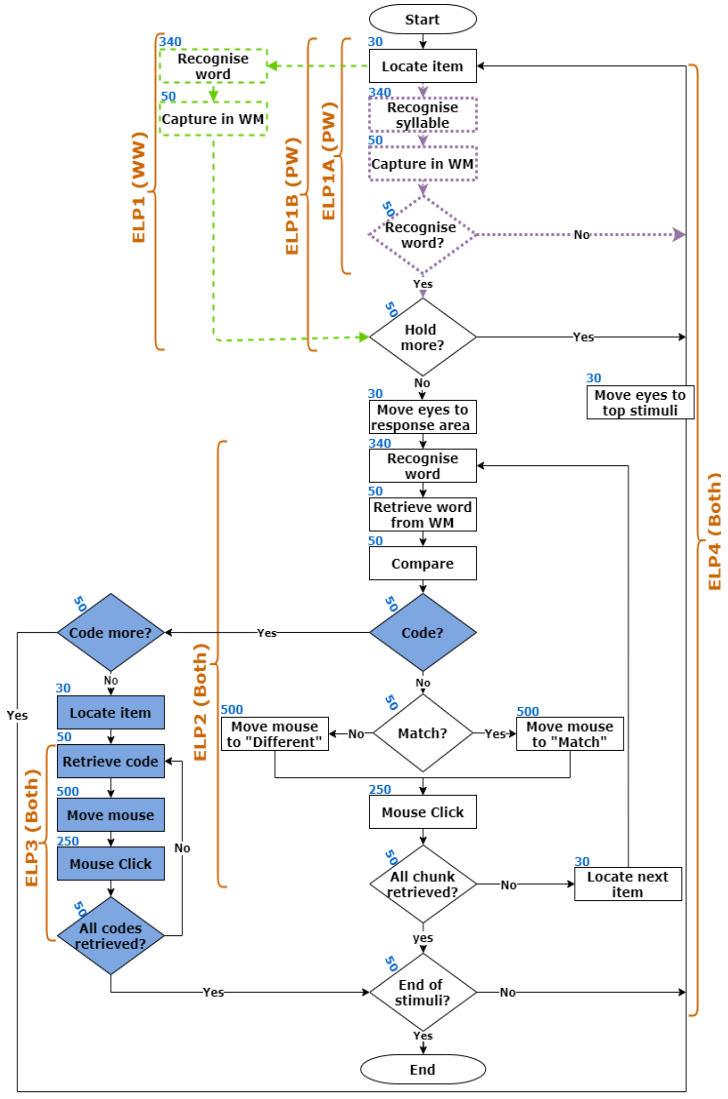

Figure 4: Expert Model

pick strategy by entering the NLP2 loop as many times as needed until certainty is attained.

In PW matching tasks, the variations in the number of stimuli views, as seen in Fig 2. P31(D,E,F) are mainly explained by two loops; NLP1(PW) and NLP2 in Fig. 3. If a novice finds it difficult to locate word boundaries, then a process of chunking one or more syllables without reaching a complete word might be applied. This means that at any point in time they might choose to opt out of NLP1(PW) and proceed to compare the chunked parts via NLP3, then loop back to the stimuli via NLP4 and continue on in this process until the whole word is compared. Another explanation, as shown earlier, might be due to uncertainty and thus entering the NLP2 loop. The NLP1(PW) and NLP2 might be experienced as many times as needed in a manner that includes either one or both of them until a response is provided. This is then reflected in their number of views and pause lengths.

According to Fig. 2, experts mainly varied amongst each other in the number of words chunked into their WM. This is due to their preference of opting out of ELP1(WW) and ELP1B(PW) in Fig. 4 at any point in time without fully loading their WM. This might be due to the nature of the task which did not put a premium on loading WM to capacity as much as possible

Finally, the recoding strategy observed in the performances in both groups can be explained by the "code" decision process (see the blue colored flow of processes in Fig. 3& 4). If the participant finishes comparing the memorized item to the response, they might choose to recode their decisions rather than clicking answer, thus viewing the stimuli multiple times, comparing and then generating a list of codes. Once enough codes have been produced, they would enter the NLP5 (Fig. 3) or ELP3 (Fig. 4) loop by simply retrieving one decision code at a time and clicking their options.

The models produce a range of pause durations based on the type and number of loops encountered. To evaluate the

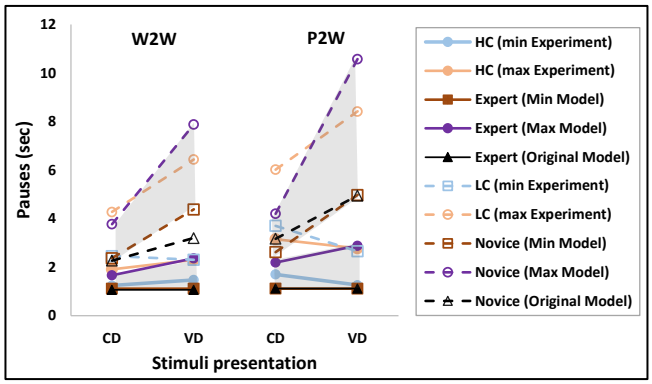

Figure 5: Model evaluation

predictions (Fig. 5), an expert's max and min pauses were calculated based on chunk size (ELP1(WW) & ELP1B(PW) in Fig. 4). By observing the HC participant's performances, their chunk size ranged between one to three words per view, and thus were chosen to represent the expert's boundaries. Consisted with the LC's performance, a novice's chunk size was limited to one word, however the number of times entering the confusion or uncertainty loops are what determined the novice's model limits (NLP1(PW) & NLP2(both) in Fig. 3). Therefore, the min value was set at no entry to those loops, while the max was based on encountering each loop once for every word processed. In Fig. 5, the solid lines indicate to the expert and HC data, while the dotted represent the novice and LC. Moreover, the black lines represent the original model, the dark point to the min and max boundaries of the new model, and the light lines represent the experimental data.

Results show that the original model was always at the lower bound of the range of participants, while the modified model encompasses much more of the range, excluding some of the LC's min values and a few of the HC's max values.

Discussion

We are developing the Competence Assessment by Stimulus Matching (CASM) technique for the assessment of competence in natural language that exploits measurements of chunk signals. A summary of an empirical evaluation of the CASM was presented and compared against the initial GOMS models (Ismail & Cheng, 2021) used to design CASM tasks. The models' predictions were partially supported. Overall, high competent individuals experienced shorter pause durations prior to clicking answers and made a smaller number of stimuli views compared to less experienced counterparts. This likely reflects the different chunk structures between the two groups, conforming to the chunking theory (Cowan, 2001; Gobet et al., 2001; Miller, 1956). Moreover, the results are in line with previous studies that used hand transcription tasks to measure competence in various domains (Albehaijan & Cheng, 2019; Cheng & Rojas-Anaya, 2007; Zulkifli, 2013).

However, the experiment revealed major intra-participant differences otherwise not captured by the initial GOMS model. To address the limitation, we examined the participants' strategies when interacting with the VD tasks. Three main observations were made.

First, low competent individuals differed amongst each other in their number of views. This might have been caused by either uncertainty of items held in WM, or inability to group the syllables into a word causing participants to loop the associated processes in the model (Fig. 3) as many times as needed until a word is recognized. Both cases are explained by an absence of chunks pertaining these words in long term memory. The weaker the chunk, the higher the chance of these loops occurring, causing an increased number of views.

Second, experts' ability to load a large number of words into WM, reflects the existence of those chunks in their long-term memory. However, some participants in this group did not fully utilize their WM, by limiting their chunking following a *single-view-single-pick* strategy. This is caused by not performing enough loops in the initial processes pertaining to word recognition and chunking (Fig. 4) The main aim of CASM is for experts and novices to be loading WM to the same extent with numbers of chunks, so that they are comparable in that regard, but what differs between them is the size of the chunks, which will be larger for the experts than the novices, hence a better performance.

Third, participants across groups applied a recoding strategy, that reflects nothing of their language ability. Following this strategy, the information contained within their chunks is a code of their potential responses rather than the words themselves. Such could assist the participants in managing their working load, as they drop information pertaining the words early on and retain a much easier to memorize code.

The evaluation results show that the new model, though not producing perfect matches, out-performs the original one. As for the out-of-range values, it was observed that the low competent individuals, in many instances, were employing the recoding strategy, which was not modeled in Fig. 5. However, there is no observed explanation for the high competent, but we hope to find out in subsequent experiments.

The revision of the GOMS models provided for a better understanding of the sources that caused the inter and intra-participant differences. This is helpful for the future refinement of the CASM task in at least two ways. First, to increase the demands of the task to encourage individuals to load up their WM, hence use their chunking ability more. Second, to eliminate the possibility of individuals applying a recoding strategy. We are modifying the design CASM tasks.

From a wider perspective, this study took an incremental step towards using GOMS to develop a model that includes various individual differences, which challenges the claim made by Olson and Olson (1990). Therefore, a particular contribution of this work is the demonstration how, in one way, GOMS models may address individual differences.

References

Albehaijan, N., & Cheng, P. C.-H. (2019). *Measuring programming competence by assessing chunk structures in a code transcription task.*

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*: Crc Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology, 4*(1), 55-81.

Cheng, P. C.-H. (2014). Copying equations to assess mathematical competence: An evaluation of pause measures using graphical protocol analysis. *Proc. of the 36th Annual Conf. of the Cognitive Science Society*, 319-324.

Cheng, P. C.-H., & Rojas-Anaya, H. (2007). *Measuring mathematic formula writing competence: An application of graphical protocol analysis.* Paper presented at the Proceedings of the Annual Meeting of the Cognitive Science Society.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behav Brain Sci, 24*(1), 87-114.

Gobet, F., Lane, P. C., Croker, S., Cheng, P. C., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in cognitive sciences, 5*(6), 236-243.

Ismail, H. B., & Cheng, P. C.-H. (2021). *Competence assessment by stimulus matching: an application of GOMS to assess chunks in memory*. Paper presented at the proceedings of the 19th ICCM Conference.

John, B. E., & Kieras, D. E. (1994). *The GOMS family of analysis techniques: Tools for design and evaluation*. Retrieved from

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological review, 101*(2), 343-352.

Olson, J. R., & Olson, G. M. (1990). The Growth of Cognitive Modeling in Human-Computer Interaction Since GOMS. *Human-computer interaction, 5*(2-3), 221-265.

Zulkifli, M. (2013). Applying pause analysis to explore cognitive processes in the copying of sentences by second language users. In: ProQuest Dissertations Publishing.